Supplementary

A. Dataset Description

As there are no datasets specifically designed for evaluating image immunization performance, we conduct quantitative experiments following settings similar to those used in Photoguard [23]. We first generate 150 images featuring 3 distinct objects using the diffusion model. For each object, we create 2 editing prompts corresponding to two malicious editing scenarios: altering specific content in the image or manipulating other regions. The details of the prompts used in the experimental dataset are provided as follows:

Original prompt	Editing prompt 1	Editing prompt 2
A dog	A cat	A dog in the park
A horse	A zebra	A horse and a cow
A man	A woman	A man with a hat

Notably, *Editing prompt 1* is designed to simulate a scenario where malicious users attempt to alter specific content in the image, while *editing prompt 2* simulates a situation where malicious users try to manipulate regions other than the specific content of our concern. Due to the variability in outcome quality caused by different random seeds [13], we conduct our experiments by averaging the editing results over 20 random seeds. The editing results are obtained using Stable Diffusion V1.4 [19] in accordance with the provided editing prompts.

B. Attacking Strength vs Diffusion Timestep

The strength of our semantic attack is tied to the number of diffusion timesteps employed in the immunization. As illustrated in Figure A, it is evident that similar to the diffusion attack [23], our semantic attack becomes more potent with an increase in the number of attacked diffusion timesteps. Notably, our approach is able to achieve stronger immunization under the same number of diffusion timesteps while reducing GPU memory usage (see Figure 6), making it more computationally efficient.

C. The Effect of Semantic Attack on Attention Map

Figure B illustrates the attention decay during our semantic attack. The primary objective of our semantic attack is to divert the attention of the diffusion model, making it unable to locate the correct region for editing. As depicted in the figure, after a few iterations of attacking, the attention of the model is effectively ”distracted,” and the attention map of the concerned content is sabotaged, leading to suboptimal results in subsequent editing attempts.

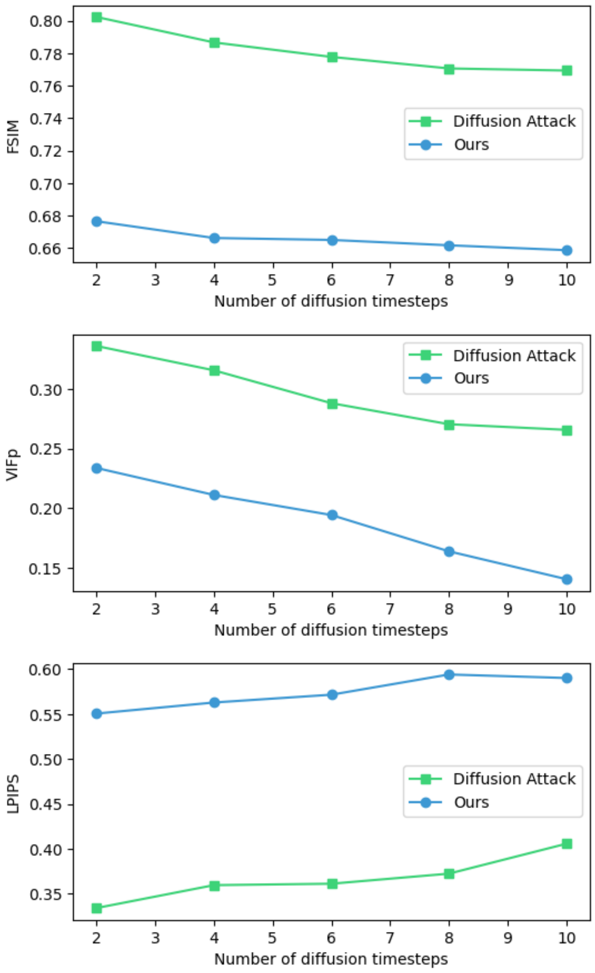

Figure A. Comparison of the quality of edited images immunized by the diffusion attack and our semantic attack under various diffusion timesteps.

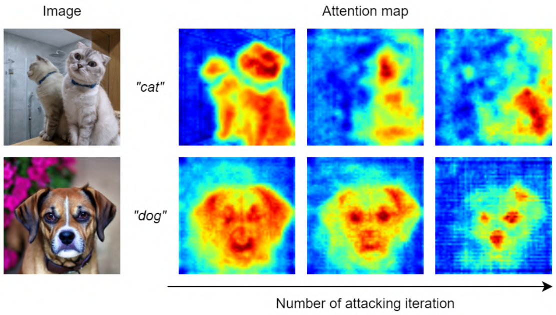

Figure B. An illustration of the attention map after .

D. Additional Qualitative Results

We present supplementary results of our semantic attack to highlight and demonstrate the immunization capabilities of our approach.

As illustrated Figure C, previous immunization approaches exhibit varying degrees of success in protecting images, with uncertainty in their effectiveness. For instance, in the figure, the immunization of certain examples, such as the polar bear image, is relatively successful, while others fail to render the editing ineffective. In contrast, our semantic attack ensures that the model cannot recognize the content of our concern while masking out other regions, providing a more stable and consistent level of protection. Through our immunization, the model struggles to generate details according to the prompt, as evident in the case of ”kitchen” and ”cowboy” in the second example. Additionally, it fails to recognize preserved content, resulting in overlapping results of two different objects, as observed in the example of the chicken. We provide additional examples of the immunization ability against image-to-image editing in Figure D. It is evident that with the same perturbation budget and attacking iterations, our approach achieves superior immunization, resulting in editing outcomes that are more unreal and exhibit more artifacts.

Additional examples in Figure E F G are provided to assess the immunization effectiveness of our semantic attack against advanced editing approaches. As highlighted in the main paper, the editing outcomes undergo significant disruption following immunization through our semantic attack.The effectiveness of Null-text inversion [14] results is significantly diminished due to their dependency on precise attention maps for the editing object. In contrast, EDICT [30] demonstrates a nullifying impact on the editing of immunized images, possibly due to its pursuit of an exact inversion of a real image using the noises predicted by the denoising U-net. Our attack on the denoising U-net may lead to a failed inversion, compromising the effectiveness of the forward diffusion process in editing. On the other hand, DiffEdit [4] outcomes exhibit a failure to generate accurate results after immunization. This can be attributed to their dependence on distinguishing the predicted noise on the original object from the desired editing object. Our model interferes with the denoising U-net, hindering the model's accurate recognition of the region to be edited, leading to low-quality editing results.

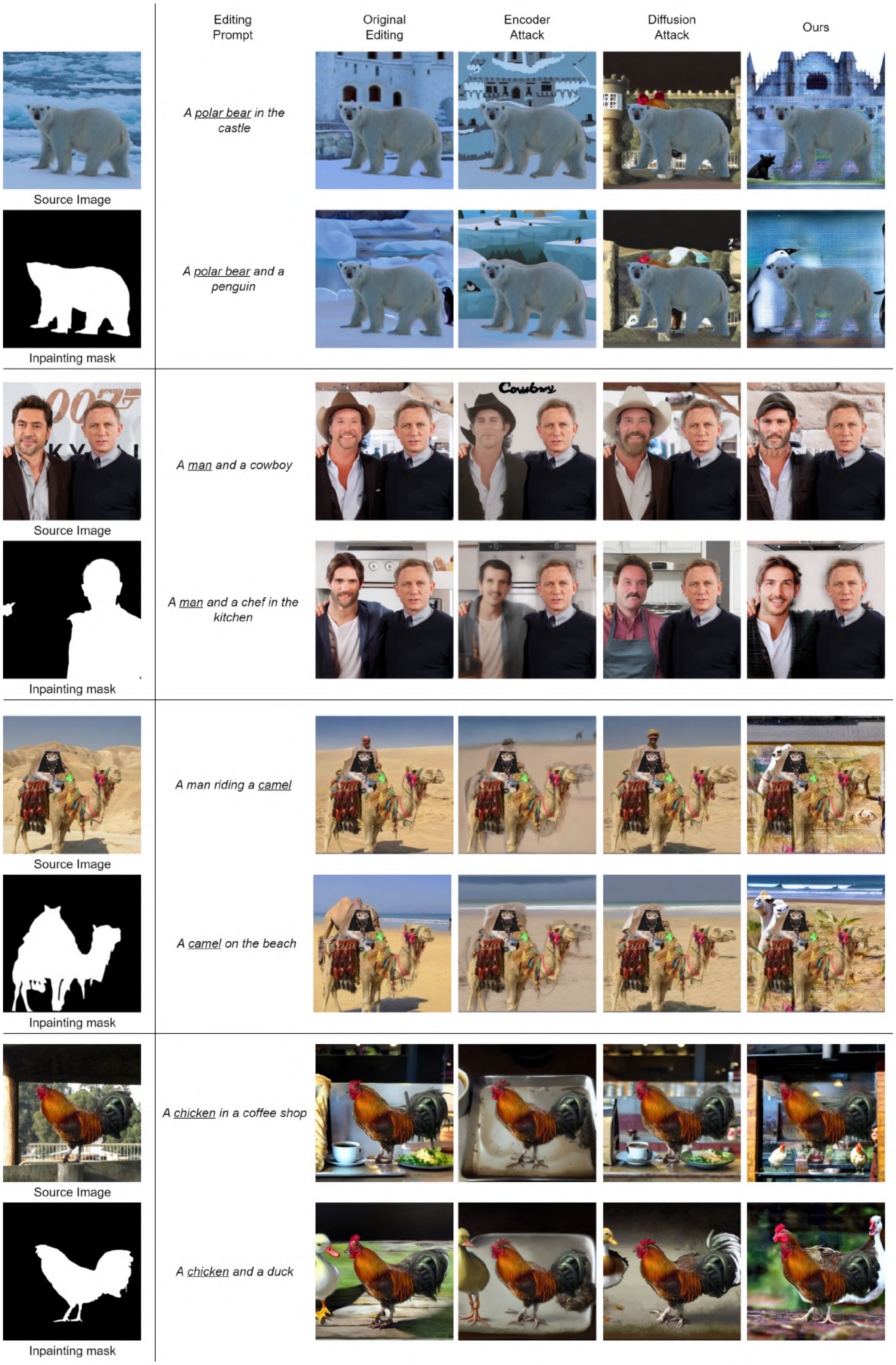

Figure C. Qualitative comparison on the immunization ability against image inpainting. The content to be immunized in our semantic attack is indicated by the underlined word in the text prompt.

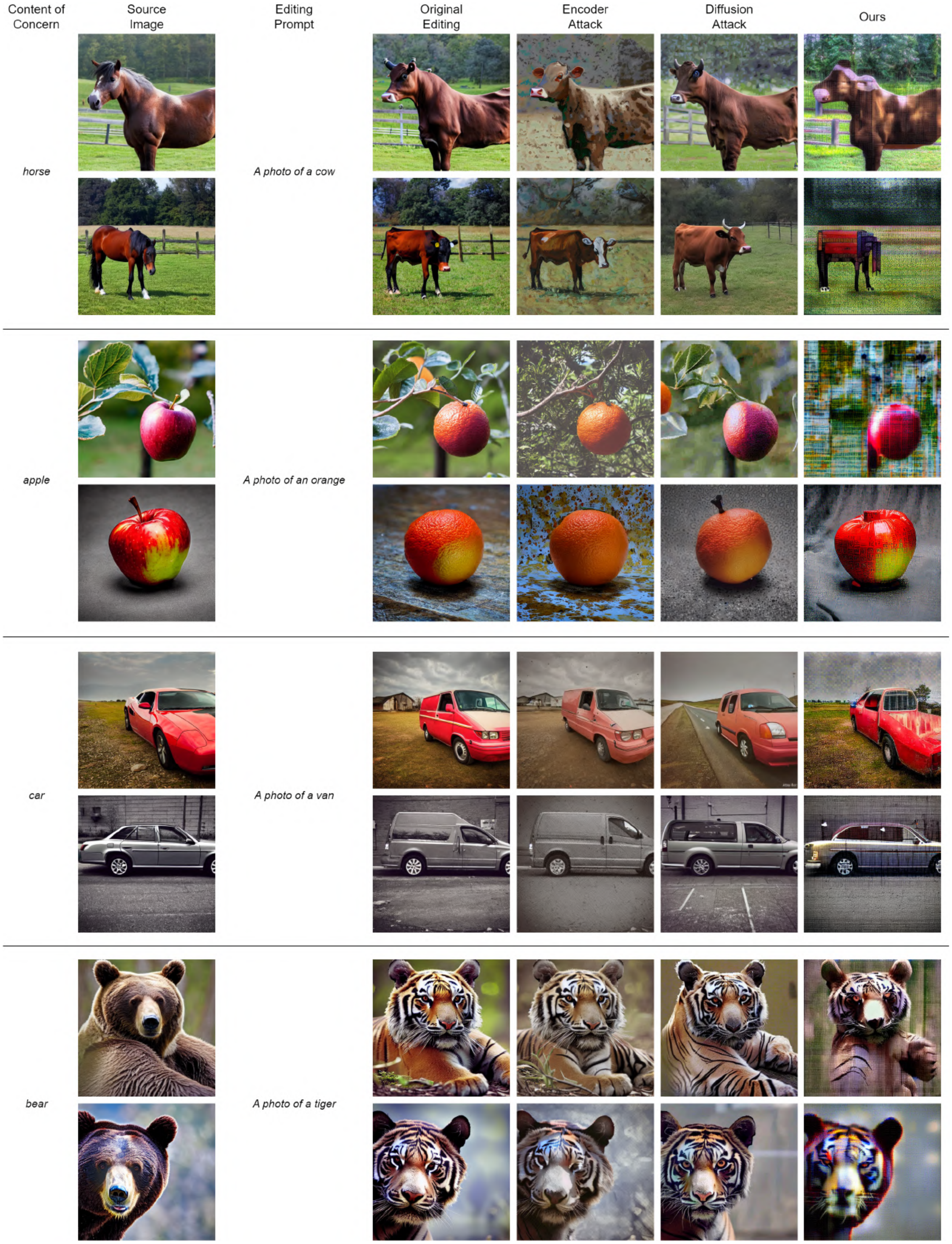

Figure D. Qualitative comparison on the immunization ability against image editing.

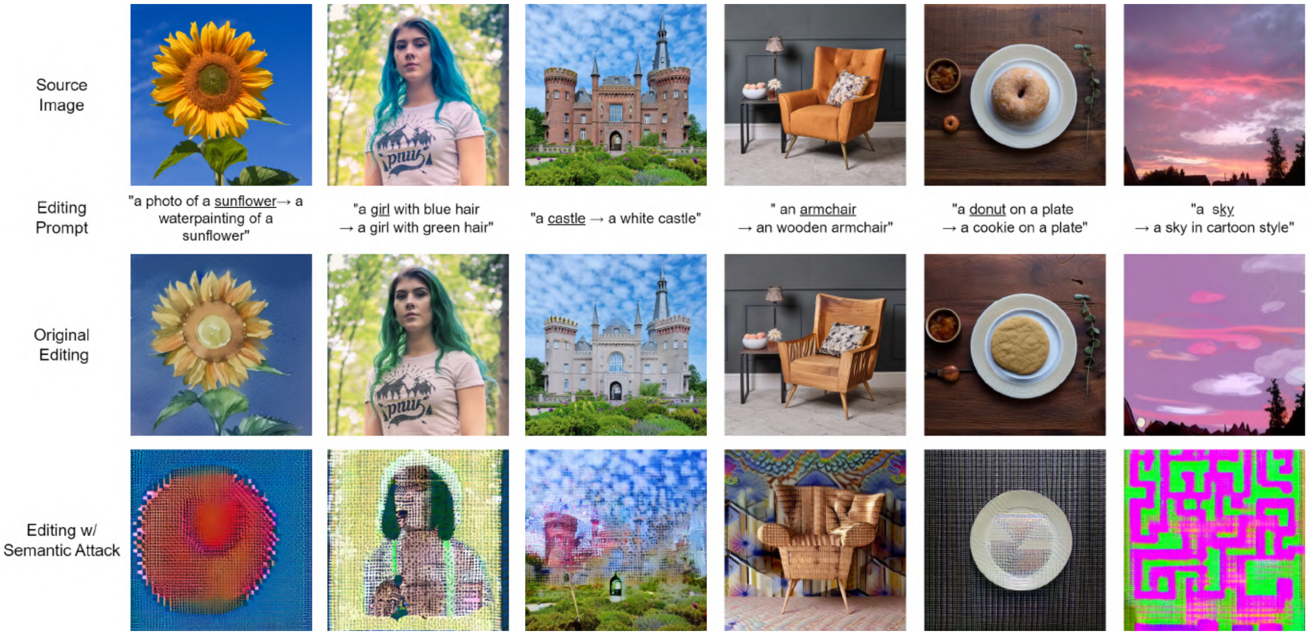
Source Image
Editing Prompt
"a photo of a sunflower→ a waterpainting of a sunflower"
"a girl with blue hair → a girl with green hair"
"a castle → a white castle"
" an armchair → an wooden armchair"
"a donut on a plate → a cookie on a plate"
"a sky → a sky in cartoon style"
Original Editing
Editing w/ Semantic Attack

Figure E. Qualitative results of our immunization against null-text inversion editing proposed in [14]. The content to be immunized in our semantic attack is indicated by the underlined word in the text prompt.

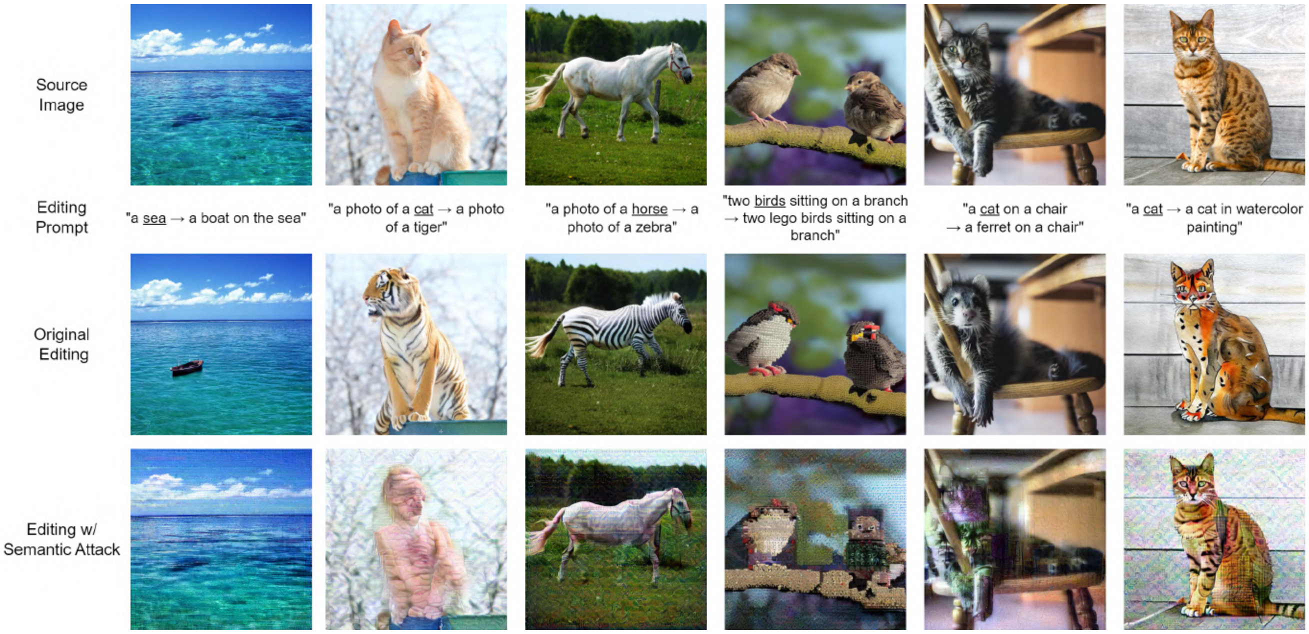
Source Image
Editing Prompt
"a sea → a boat on the sea"
"a photo of a cat → a photo of a tiger"
"a photo of a horse → a photo of a zebra"
"two birds sitting on a branch → two lego birds sitting on a branch"
"a cat on a chair → a ferret on a chair"
"a cat → a cat in watercolor painting"
Original Editing
Editing w/ Semantic Attack

Figure F. Qualitative results of our immunization against EDICT editing proposed in [30]. The content to be immunized in our semantic attack is indicated by the underlined word in the text prompt.

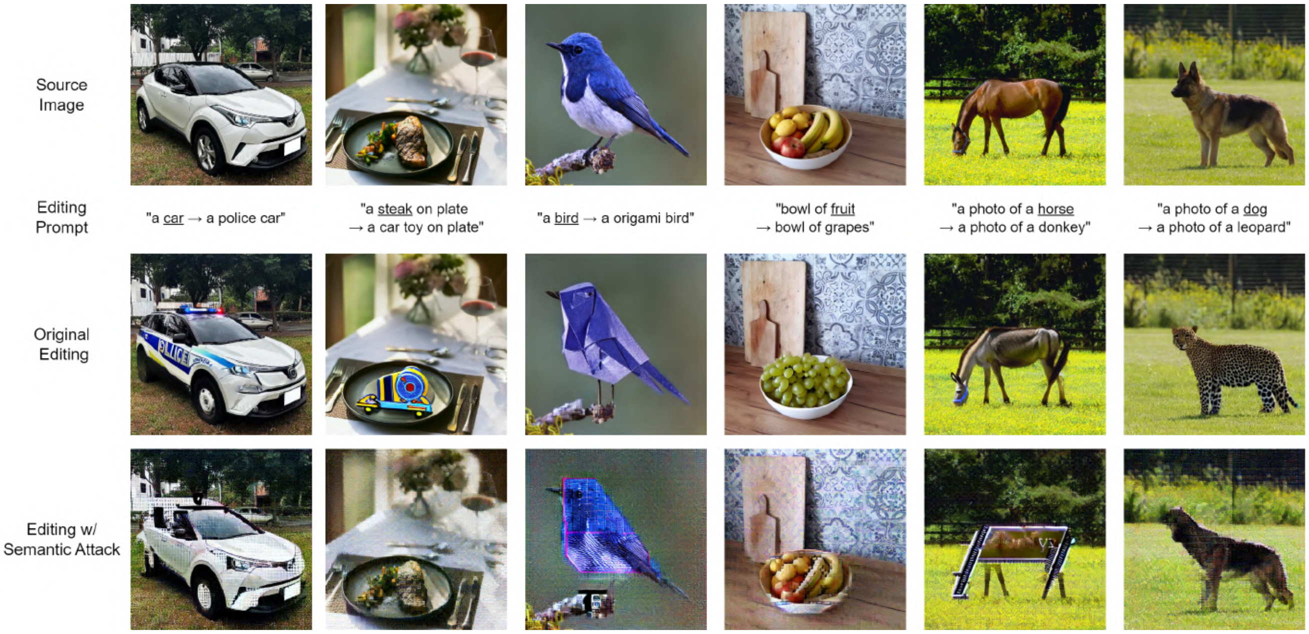
Source Image
Editing Prompt
"a car → a police car"
"a steak on plate → a car toy on plate"
"a bird → a origami bird"
"bowl of fruit → bowl of grapes"
"a photo of a horse → a photo of a donkey"
"a photo of a dog → a photo of a leopard"
Original Editing
Editing w/ Semantic Attack

Figure G. Qualitative results of our immunization against DiffEdit editing proposed in [4]. The content to be immunized in our semantic attack is indicated by the underlined word in the text prompt.